J. Mol. Biol. (1994) 242, 309–320

REVIEW ARTICLE

The Immunoglobulin Fold

Structural Classification, Sequence Patterns and Common Core

P. Bork, L. Holm and C. Sander

EMBL, Meyerhofstraße 1, D-69012 Heidelberg, Germany

Since the first crystal structure of an immunoglobulin revealed a modular architecture, the characteristic β-sheet fold of the immunoglobulin domain has been found in many other proteins of diverse biological function. Here, a systematic comparison of 23 Ig domain structures with less than 25% pairwise residue identity was performed using automatic structural alignment and analysis of β-sheet and loop topology. Sequence consensus patterns were identified for nine distinct families with at most marginal similarity to each other. The analysis reveals a common structural core of only four β-strands (*b*, *c*, *e* and *f*), embedded in an antiparallel curled β-sheet sandwich with a total of three to five additional strands (*a*, *c'*, *c"*, *d*, *g*) and a characteristic intersheet angle. The variation in the position of the edge strands (*a*, *c'*, *c"*, *d* and *g*) relative to the common core defines four different topological subtypes that correlate with the length of the intervening sequence between strands *c* and *e*, the most variable region in sequence. The switch of strand *c'* from one sheet to the other in seven-stranded domains appears to result from short *c-e* segments, rather than being a major structural discriminator. The high degree of structural flexibility outside the common core and the extreme variability of side-chain packing inside the core do not support a protein folding pathway common to all members of the structural class. Mutation rates of immunoglobulin-like domains in different proteins vary considerably. Disulfide bridges, thought to contribute to structural stability, are not necessarily invariant in number and location within a subclass.

Keywords: structural comparison; sequence similarity; molecular evolution; protein folding

1. Introduction

The exponential growth of sequence databases and the drastic increase in published tertiary structures have revealed an increasing number of protein families with similar structure but with extremely divergent sequences (reviewed by Holm & Sander, 1994). One of the most striking examples is the emerging class of immunoglobulin-like (Ig-like) domains. The folding topology of immunoglobulin constant and variable domains has been described as a Greek key β-barrel, subclass "simple" (Richardson, 1981). The structural domains of immunoglobulin have seven to nine antiparallel β-strands forming a barrel-like shape. However, hydrogen bonds do not go around the barrel so that there are two distinct β-pleated sheets and physically the fold is a β-sandwich (Lesk & Chothia, 1982). The class of simple Greek key proteins includes a number of other proteins, e.g. superoxide dismutase, blue copper proteins, hemocyanin, but they have additional elements of secondary structure. This review aims at a classification of structures of immunoglobulin and non-immunoglobulin domains which nevertheless have the same topology as immunoglobulins, i.e. the same order and number of strands.

Recently, Ig-like domains have been reported in numerous structures of non-immunoglobulins including (1) cell surface receptors such as CD2 (Driscoll *et al.*, 1991; Jones *et al.*, 1992), CD4 (Ryu *et al.*, 1990; Wang *et al.*, 1990; Garrett *et al.*, 1993; Brady *et al.*, 1993), CD8 (Leahy *et al.*, 1992*a*), MHC/HLA (Saper *et al.*, 1991 and references therein), growth hormone receptor (de Vos *et al.*, 1992), neuroglian (Huber *et al.*, 1994), (2) matrix proteins such as tenascin (Leahy *et al.*, 1992*b*) and fibronectin (Main *et al.*, 1992), (3) intracellular regulatory proteins such as the bacterial chaperonin PapD (Holmgren & Brändén, 1989), as well as (4) enzymes such as cyclodextrin glycosyltransferase (Klein & Schulz, 1991), myosin light chain kinase (telokin; Holden *et al.*, 1992), and galactose oxidase (Ito *et al.*, 1991). These proteins have diverse functions. Based on distant sequence similarity, some of the domains were predicted to belong to the immunoglobulin superfamily in advance of structure determination (e.g. type III repeat of fibronectin (Fn3†; Bazan, 1990), in other cases the structural similarity was quite unexpected (e.g. PapD; Holmgren & Brändén, 1989). All Ig-like domains appear to

† Abbreviations used: Fn3, fibronectin type III; PDB, protein data bank; r.m.s., root-mean-square; GHR, growth hormone receptor.

0022-2836/94/390309-12 $08.00/0

be involved in binding functions; not a single one is known to contain a natural enzyme active site. Considering the vast number of sequence relatives, the Ig-like domain is probably the most widespread protein module, at least in animals (Doolittle & Bork, 1993). Based on sequence information, 40% of the human leucocyte surface proteins were predicted to contain Ig-like domains (Barclay *et al.*, 1992); this number might even increase considering the various sequence-unrelated proteins and their families reviewed here.

Faced with the increasing functional, structural and sequence diversity of Ig-like domains, one wonders whether there are any conserved features common to all Ig-like domains. If there are, do they indicate a common folding principle or common folding pathway? Is it possible to discriminate divergent evolution within this family from convergence towards an energetically stable fold? What is the correlation between sequence similarity and structural similarity in marginally related sequences? To provide a basis for answering these questions we have undertaken an objective (automatic) multiple superimposition of all these domains and correlated the common and distinct features with functional and sequence information.

Table 1

Classification of structures used in the analysis

Code	Protein	Species	Topology	Res. (Å)	s – s (deg)	aa c^*-e	Refs
2CD4-2	CD4 domain 2	*Homo sapiens*	s-type	2·3	153	7	Ryu *et al.* (1990)
2HHR-2	Growth hormone receptor, domain 2	*Homo sapiens*	s-type	2·8	147	7	deVos *et al.* (1992)
1TEN	3rd fibronectin-type repeat of tenascin	*Homo sapiens*	s-type	1·8	153	8	Leahy *et al.* (1992*b*)
1CID-4	CD4 domain 4	*Rattus rattus*	s-type	2·8	154	9	Brady *et al.* (1993)
CD2-2	CD2 domain 2	*Rattus rattus*	s-type	2·8	160	9	Jones *et al.* (1992)
FN3	Fibronectin	*Homo sapiens*	s-type	NMR	150	9	Main *et al.* (1992)
GN-1	Neuroglian, domain1	*Drosophila melanogaster*	s-type	2·0	152	9	Huber *et al.* (1994)
GN-2	Neuroglian, domain 2	*Drosophila melanogaster*	s-type	2·0	141	10	Huber *et al.* (1994)
1GOF	Fungal galactose oxidase, domain 3, res. 533–639	*Dactylium dendroides*	h-type	1·7	150	12	Ito *et al.* (1991)
1CGT	Cyclodextrin glycosyl-transferase domain D, res. 495–580	*Bacillus circulans*	h-type	2·0	n.d.‡	13	Klein & Schulz (1991)
2HHR-1	Growth hormone receptor, Domain 1	*Homo sapiens*	s-type	2·8	159	16	deVos *et al.* (1992)
1TLK	Telokin	*Meleagris gallopavo*	c/v-type†	2·8	160	17	Holden *et al.* (1992)
3DPA-1	papD protein, domain 1, res. 1–120	*Escherichia coli*	s-type	2·5	141	17	Holmgren & Branden (1989)
3-FAB-VL	IgG2a-kappa, variable domain of light chain	*Muso musculus*	v-type§	2·0	154	18	Herron *et al.* (1989)
3FAB-CH	IgG2a-kappa, constant domain of heavy chain	*Muso musculus*	c-type	2.0	162	20	Herron *et al.* (1989)
CEL	Cellulase CelD	*Clostridium thermocellum*	h-type	2·3	141	21	Juy *et al.* (1992)
1FC1-C2	Constant domain 2	*Homo sapiens*	c/v-type†	2·9	160	21	Deisenhofer (1981)
1FC1-C3	Constant domain 3	*Homo sapiens*	c-type	2·9	157	21	Deisenhofer (1981)
3HLA	Class I histocompatib. antigen A2·1	*Homo sapiens*	c-type	2·6	165	21	Saper *et al.* (1991)
2HLA	Class I histocompatib. antigen A68·1	*Homo sapiens*	c-type	2·6	165	22	Garrett *et al.* (1989)
CD2-1	CD2 domain 1	*Rattus rattus*	v-type§	2·8	168	23	Jones *et al.* (1992)
2RHE-VL	Bence-Jones Ig variable domain	*Homo sapiens*	v-type§	1·6	152	25	Furey *et al.* (1983)
1CD8-1	CD8 domain 1	*Homo sapiens*	v-type	2·6	150	29	Leahy *et al.* (1992*a*)
1CID-3	CD4 domain 3	*Rattus rattus*	v-type§	2·8	157	29	Brady *et al.* (1993)
2CD4-1	CD4 domain 2	*Homo sapiens*	v-type§	2·3	151	32	Ryu *et al.* (1990)
4FAB-VH	IgG2a-kappa, variable domain of heavy chain	*Homo sapiens*	v-type	2·0	145	34	Herron *et al.* (1989)

† Telokin has 8 strands and is structurally similar to v-type but lacks c^*. Thus, it appears to be a hybrid between c-type and v-type. The constant domain 2 of 1FC1 is most similar to c-type structures but has a very short strand topologically equivalent to c'.

‡ Not determined due to a β bulge involving the residues used to calculate the sheet-sheet angle.

§ Strand a switched over to sheet g-f-c-c'-c^*.

Proteins studied are identified by their PDB code, if available, followed by a hyphen and domain mnemonic. The co-ordinates for CD2. FN3, neuroglian and cellulase CelD were kindly provided by the authors. The structures are ordered according to their loop lengths between the reference point c^* and e. Note the correlation between the loop length and the structural subtype. There are 4 loops between the 5 segments of the structural core (b-c, c-c^*, c^*-e, e-f). The number of residues in the c^*-e segment is given in column (aa c^*-e). Sheet-sheet angles (column s-s deg) were defined as the dihedral angle between vectors based at the midpoint of strands b and c and of strands e and f, counting the residues of the common core only. The orientation of the vectors was the sum of all CA_i to CA_{i+1} vectors for strand b and CA_i to CA_{i-1} vectors for strand e, and similarly for strands c and f.

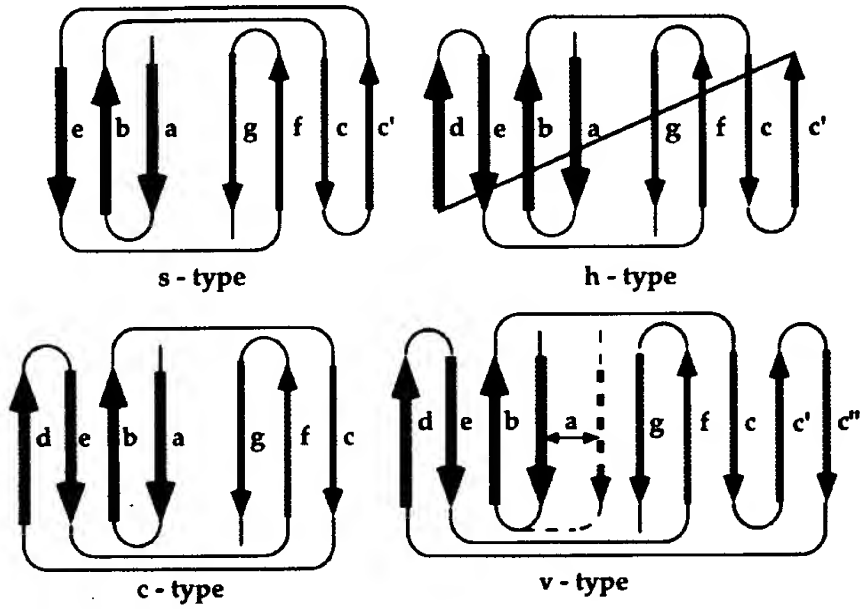

Figure 1. 2D topology diagrams of observed hydrogen bonding patterns. The 7–9 strands (*a*, *b*, *c*, *c'* *c"*, *d*, *e*, *f*, *g*) form a sandwich of 2 sheets (back sheet, I, thin arrows; front sheet, II, thick arrows, where back and front refer to Figure 3; packing not evident and loop lengths not to scale in this projection). The common core (Figure 3) is shown in red. Immunoglobulin constant domains have 7 strands in a topology shown at the bottom left (c-type, for constant). Immunoglobulin variable domains have an additional hairpin (*c'*-*c"*) between strands *c* and *d*, with a total of 9 strands (v-type, for variable). Strand *a* has 2 alternative locations in v-type domains, being part of either the back sheet (antiparallel pairing with strand *b*) or of the front sheet (parallel pairing with strand *g*). Other Ig-like domains also have 7 strands, but are different from c-type in that the 4th strand has switched sheets (s-type, for switched); the name of the 4th strand changes from *d* to *c'* to reflect the sheet switch. The last type represents an 8-stranded hybrid between c- and s-type that has both strands *c'* (front sheet) and its direct continuation strand *d* (back sheet), so that both sheets have 4 strands (h-type, for hybrid).

2. Structure Comparison

(a) *Search for 3D structures closely related to Igs*

Structure comparisons were carried out using the program Dali, which maximizes a geometrical similarity score calculated from intramolecular distances in the common core (details in Holm & Sander, 1993). Structures similar to immunoglobulin domains were identified from an all-against-all comparison of a representative set of protein structures (Holm & Sander, 1993). This set was extended by Ig-like structures not yet in the Protein Data Bank (PDB; Bernstein *et al.*, 1977). In addition to immunoglobulins, the selected set includes domains from HLA, CD2, CD4, CD8, fibronectin, tenascin, neuroglian, the growth hormone receptor, bacterial domains from a thermostable cellulase, cyclodextrin glycosyltransferase, PapD and a fungal galactose oxidase (Table 1). Only three pairs in this set have sequence identities at or above 25% (FN3 with 1TEN, 3FAB heavy constant domain with 1FC1 third domain, and the 2RHE and 3FAB light chain variable domains). In spite of the striking similarities between the first domain of PapD and Ig-like domains, the second domain of PapD is a complete outlier. It has very low structure similarity when compared with all other domains in the Ig-like set, and was therefore excluded. Other proteins structurally most similar to the Ig-like set include blue copper proteins, actinoxanthin and superoxide dismutase. However, these were found to be topologically distinct (additional strands inserted between *a* and *b* in several blue copper proteins; an extra strand before strand *a* in superoxide dismutase) and geometrically different (poorer superimposition, different twist and tilt angles between the sheets) and were therefore excluded from the detailed comparisons.

(b) *Definition of topological subtypes*

Current classification schemes of classical Ig-like domains are mainly based on the number of strands and sequence similarity, (division into v, c1 and c2 sequence "sets"; Williams & Barclay, 1988; Hsu & Steiner, 1992). However, with the crystal structures of PapD and CD4 (Holmgren & Brändén, 1989; Ryu *et al.*, 1990; Wang *et al.*, 1990) it became obvious that the *d*-strand of the seven-stranded structures can switch between the two sheets (then called *c'*; Figure 1) and many authors distinguish between the two forms. Backbone hydrogen bonding patterns (Kabsch & Sander, 1983) in the Ig-like domains indeed define two sheets rather than a closed barrel. The edges of the sheets are conformationally flexible (in particular, strands *a*, *g*, *c'*, *c"*). In a number of cases a strand starts

in one sheet, then the chain bulges a bit but continues in the same direction ending as a strand in the second sheet. Most recently, comparison of telokin (Holden *et al.*, 1992) with other immunoglobulin-like domains such as CD4 and CD8 lead to the proposal of a distinct "I-set" which can be detected at the sequence level by profile searches (Harpaz & Chothia, 1994).

Based on the number of strands and the location of strand *c'/d* we can define at least four distinct subtypes (Figure 1) that are present in our dataset of superimposed immunoglobulin-like domains: (1) c-type: classical seven-stranded topology of the constant domains in immunoglobulins (sheet I: *d-e-b-a*; sheet II: *g-f-c*); (2) s-type: seven-stranded strand-switched type (sheet I: *e-b-a*, sheet II: *g-f-c-c'*); (3) h-type: a hybrid between (1) and (2), where strand *c'/d* is kinked and the N-terminal residues of this segment form hydrogen bonds with strand c (sheet II) whereas the C-terminal residues belong to sheet I, and (4) v-type: a nine-stranded type as occurring in the variable domains of immunoglobulins (sheet I: *d-e-b-a*; sheet II/ *g-f-c-c'-c"*). However, the location of strand *a* within sheet I or II varies and in the future a further classification may be applied to distinguish between the forms with sheets composed of 4 + 5 or 3 + 6 strands. The length of strands *c'* and *c"* varies. In some cases, strand *c"* is just a loop and is beyond the threshold for detection by standard programs for secondary structure definition such as DSSP (Kabsch & Sander, 1983).

Although we found no example in our dataset, in principle a switch of strand *a* in the seven-stranded sandwich (similar to that seen in nine-stranded sandwiches) should also be possible. Thus, we expect a variety of additional subtypes for which 3D structures are not yet available.

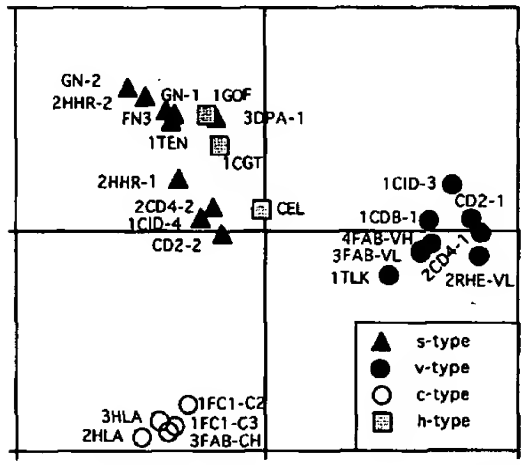

Figure 2. Clusters in structure space. A multivariate statistical analysis method called correspondence analysis (Hill, 1973) was used to represent all pairwise similarities between Ig-like domains in 2 dimensions. The distances in the plane approximately indicate structural dissimilarity. Adjacent points are most related in 3D structure. Structures are labeled as in Table 1. The clustering into subtypes and the resulting classification into v-, c- and s-type is a straightforward consequence of the structural alignment method, without predefined notions of topological types. The 1st eigenvector (horizontal) discriminates between the 9-stranded v-type and the 7-stranded domains. The 2nd eigenvector (vertical) discriminates between the 7-stranded c-type and the 7-stranded strand-switched s-type. In a further dimension, all domains from enzymes (galactose oxidase, 1GOF; cyclodextrin glycosyltransferase, 1CGT; and cellulase, CEL) form a separate group of h-type domains. Some subfamilies become visible such as CD4 domain 4 with CD2 domain 2 which form a distinct subgroup of s-type domains.

(c) *Correlation between topological subtype and structural similarity*

Multivariate analysis of the pairwise structural similarity scores (Figure 2) reveals three or four main clusters although the proteins represent a larger number of sequence families (see below). The clusters in structure space correspond to the topological classes. Projection onto the first two eigenvectors separates clusters of c, v and s(h)-type domains (Figure 2). The third eigenvector separates the first domain of PapD as an outlier; the fourth eigenvector separates the Ig-like domains surrounding the catalytic domains of phylogenetically old enzymes (i.e. in cellulase, cyclodextrin glycosyltransferase and galactose oxidase) from the s-type cluster (data not shown). Closer inspection of the clusters reveals some interesting classifications: domain 2 of the human growth hormone receptor, not only domain 1, is pulled towards the s-type Fn3 cloud; the T-cell receptor domains from CD2 and CD4 are surprisingly similar; and CD8 clusters among v-type domains (Figure 2). Telokin, an eight-stranded structure, clearly clusters within the v-type family although it lacks the *c"* strand. Average linkage clustering reveals a relatively close resemblance between the groups of the seven-stranded c-type and the nine-stranded v-type domains (not shown). These observations indicate that the transformation between seven and nine-stranded topology only involves a local perturbation of structure.

(d) *Common structural core*

We have emphasized the structural flexibility and the presence of structural subtypes in the set of Ig-like domains under study. The structural alignments allow the definition of a common structural core as those residues which are aligned against a reference structure (e.g. the variable domain of Rhe) in all structures. The common core contained strands *b*, *c*, *e* and *f* plus a piece of strand *c'* or a piece of the *c-d* loop in classes that lack strand *c'*, a total of 31 residues (Figure 3). The structural core does not contain the conformationally flexible (between families) edge strands. The reason for this is that the non-central strands can be shifted structurally between different pairs of domains. The lengths of the strands and surrounding loop regions vary extremely (Table 1).

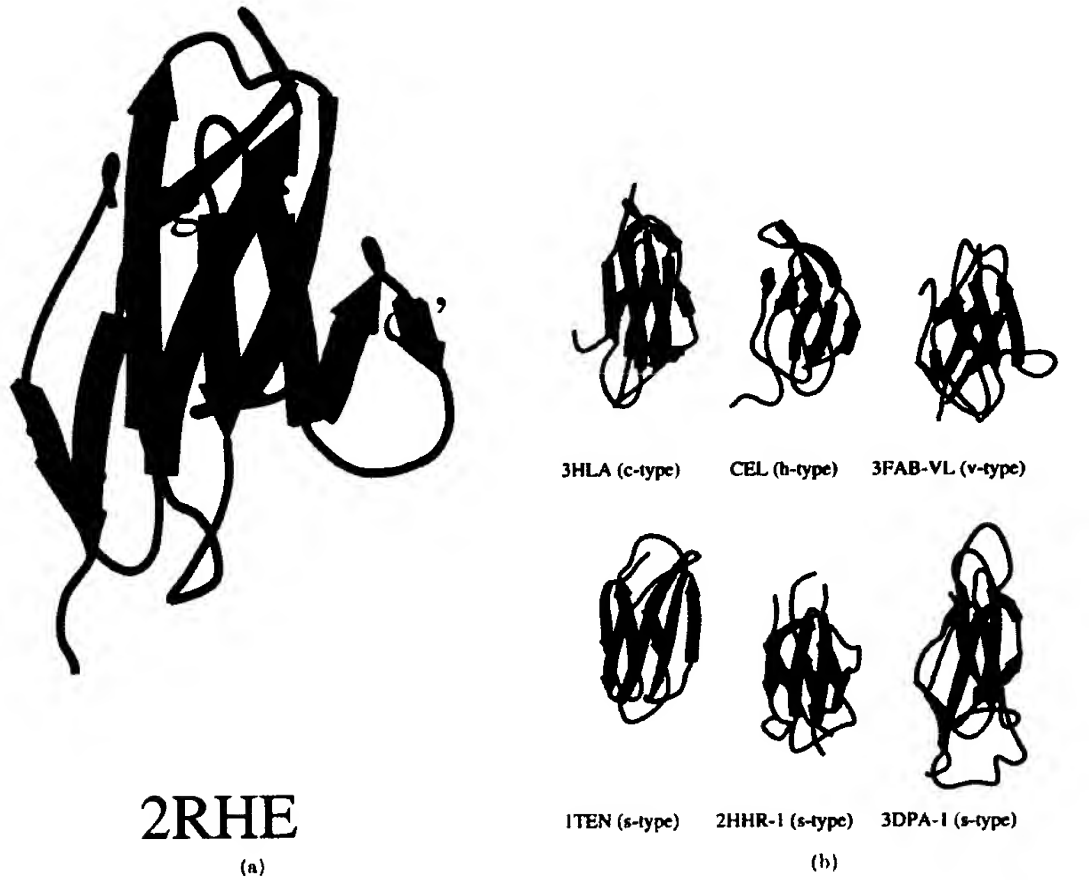

Figure 3. The common core of Ig-like domains. (a) The 2-plus-2 stranded structural core (strands *b*, *c*, *e*, *f*) common to all Ig-like domains (red) is surrounded by structurally more variable strands (green). The front sheet has up to 5 strands (*g*-*f*-*c*-*c'*-*c"*), the back sheet up to 4 (*a*-*b*-*e*-*d*). Strand *c"* is very flexible in Ig variable domains and does not always form β-strand type hydrogen bonds. The common core was defined using a multiple structural alignment generated by the program Dali (details in Holm & Sander, 1993). Ribbon by Molscript (Kraulis, 1991). Selected examples from different subtypes are shown in (b). Note that strand definitions according to DSSP do not always correspond to structurally equivalent positions.

(e) *Correlation between loop length and sheet switching*

The polypeptide chain segment between strands *c* and *e* has to change direction by at least 180 degrees and cross over to the other sheet. There is a correlation between the number of residues between strand *c* and *e* and the topological subtype. Our structural comparison revealed a reference point at the beginning of strand *c'* or in the *c*-*d* loop, respectively (Figure 1, Table 1) that could be superimposed in all pairs (short red loop segment preceding strand *c'* in Figure 3). The following observations consider the sequences between this reference point, hereafter called *c**, and the beginning of strand *e*.

In the structures of all Fn3 domains as well as the second domains of CD2, CD4 and the first domain of PapD, strand *c'* is hydrogen bonded to sheet I (thick in Figure 1, see also Figure 3). These structures (s-type) have only seven to ten residues between *c** and *e*, compared to typically well over 20 residues in c-type domains. The hybrid-type domains (intermediates between s- and c-type) of galactose oxidase and cyclodextrin glycosyltransferase have 12-13 residues between *c** and *e*. The topologies of c-type and v-type are very similar if the two additional strands *c'* and *c"* of v-types are treated as a long insertion between *c* and *e* that forms a stable β-hairpin. Following this idea, telokin (Holden *et al.*, 1992) is an intermediate between c- and v-type: 16 residues between *c** and *e* are not sufficient for a *c"* strand. Thus we see a tendency for the segment between *c** and *e* to prefer backbone-hydrogen-bonded structure over coil, folding into a β-strand or hairpin which is appended to the sheet nearest at hand. The correlation is illustrated in Table 1: the length of the intervening sequence between *c** and the *e* strand increases from s- to h- to c- to v-type structures.

3. Sequence Comparison

(a) *Sequence database searches*

The proteins of known 3D structure reviewed here represent seven distinct sequence families (considering the immunoglobulin subtypes as being one family). Two sequence families, which have been predicted to contain the Ig-fold, namely interferon receptors and CUB domains (Bazan, 1990; Bork & Beckmann, 1993) were added although their structures have not yet been solved. Figure 4 shows the derived conserved features of each sequence family. The number of sequences that are available in current sequence databases varies from one family to another (Figure 4). The most abundant families include the domains of immunoglobulins. More than 1000 members of the Igc family are already stored in current sequence databases. Nevertheless, the consensus (determined at a 70% conservation level) clearly shows conserved hydrophobic features (Figure 4).

(b) *Comparison of consensus sequences of the families*

The presence of a conserved structural core does not imply a similar hydrophobicity pattern among the different sequence families. Only strand *f* appears to retain clear conserved hydrophobic features in all sequence families. It remains unclear whether this requirement is needed to initiate or stabilize folding of the β-sandwich. Turn formation has been proposed to be a key step in folding as an initiator of the zipping up of β ladder structures. Inspection of the loops between the sequence families revealed not a single loop conserved in length throughout the families that could specifically initiate folding. Other features, such as conserved residue properties in loop segments, could not be detected either. Even within sequence families loops are not necessarily conserved and can vary in length and amino acid composition.

In most of the sequence families shown in Figure 4, aromatic residues are conserved near the termini of the core β-strands, although they do not correspond to topologically equivalent strands. Nevertheless, the presence of aromatic residues might be required to form a stable hydrophobic core. Aromatic residues have been proposed to play a special role in the design of antiparallel β-sandwiches. Their large size and specific geometry make them the best candidates to fill some specific positions, and because of their low conformational entropy they have been suggested to possibly play a role as folding nuclei (Finkelstein & Nakamura, 1993). Aromatic residues are strongly conserved within families but not between families. This conservation is particularly strong for families that lack stabilizing disulfide bridges.

(c) *Disulfide bridges*

One of the original hallmarks of an Ig-like fold was a conserved disulfide bridge between strands *b* and *f* (Lesk & Chothia, 1982). It is, however, not necessary to determine the Ig-fold. Even within the Ig-like sequence family these cysteine residues might be completely absent (see Williams & Barclay, 1988; and references therein) as, for example, shown for domain 1 of CD2 and domain 3 of CD4 (Jones *et al.*, 1992; Brady *et al.*, 1993). In other domains, the disulfide bridges have moved within the core. For example, domain 2 of CD4 forms a disulfide bond between strand *c* and *f* (Ryu *et al.*, 1990; Wang *et al.*, 1990) instead of between strand *b* and *f* as in classical Ig domains; the second domain of CD2 has in addition to the *b*-*f* bond a disulfide bridge between strands *a* and *g* (Jones *et al.*, 1992).

The number and location of disulfide bridges also varies in families without sequence similarity to Igs. The first domain of growth hormone receptor contains three disulfide bridges, other members of the homeopoietic receptor family (C4 in Figure 4; Bazan, 1990) only two. Alignment of all these sequence-

Figure 4. Comparison of distinct sequence families with Ig-like topology. The consensus sequence of each family is represented by a string of symbols (capitals, conserved amino acids; lower case, h is hydrophobic; t is turn-like or polar; a is aromatic; o is OH group = S/T; + / − is charged). Positions without any obvious consensus are indicated by a dot, .; do *not* interpret dots as gaps. Gaps are denoted by underscore, (_), surrounded by numbers that represent the variability of loop lengths within 1 sequence family. In some families, the terminal strands are too variable for a consensus to be derived; these parts have been omitted. The numbers at the right (occ, for occupancy) are estimates of the number of family members in the current protein databases. Bullets above the sequences mark structurally equivalent core positions (residues in the common core that make major inter-sheet contacts). Lower case letter strings below the sequences are strand labels, as in Figures 1 and 3. Database searches were carried out using FASTA (Pearson & Lipman, 1988). The identified proteins or domains were aligned and both profiles (Gribskov *et al*, 1987) and property patterns (Rohde & Bork, 1993) were constructed. The procedure was conducted by iteratively adding new identified members to the multiple alignment (for details, see Bork, 1993). An exception was made for the vast number of proteins sequence-related to immunoglobulins. Here, the classification of Williams & Barclay (1988) was used and 4 subfamilies were defined: Igc1, Igc2, Igv and domains without disulfide bridges. Even if constant and variable domains show sequence similarity to each other they can be well distinguished because of their different number of amino acids between strand *c* and *e*. The searching scheme resulted in a distinction of 7 sequence families of known structure which could not be fused by either sequence profile or property pattern methods. In addition to families with known structure, the sequence analysis procedure was applied to interferon receptors and CUB domains, for which an Ig-like topology has been proposed (Bazan, 1990; Bork & Beckmann, 1993). From the resulting alignment the consensus lines were derived; amino acids or properties that are conserved in more than 70% of the sequences of a particular family are displayed.

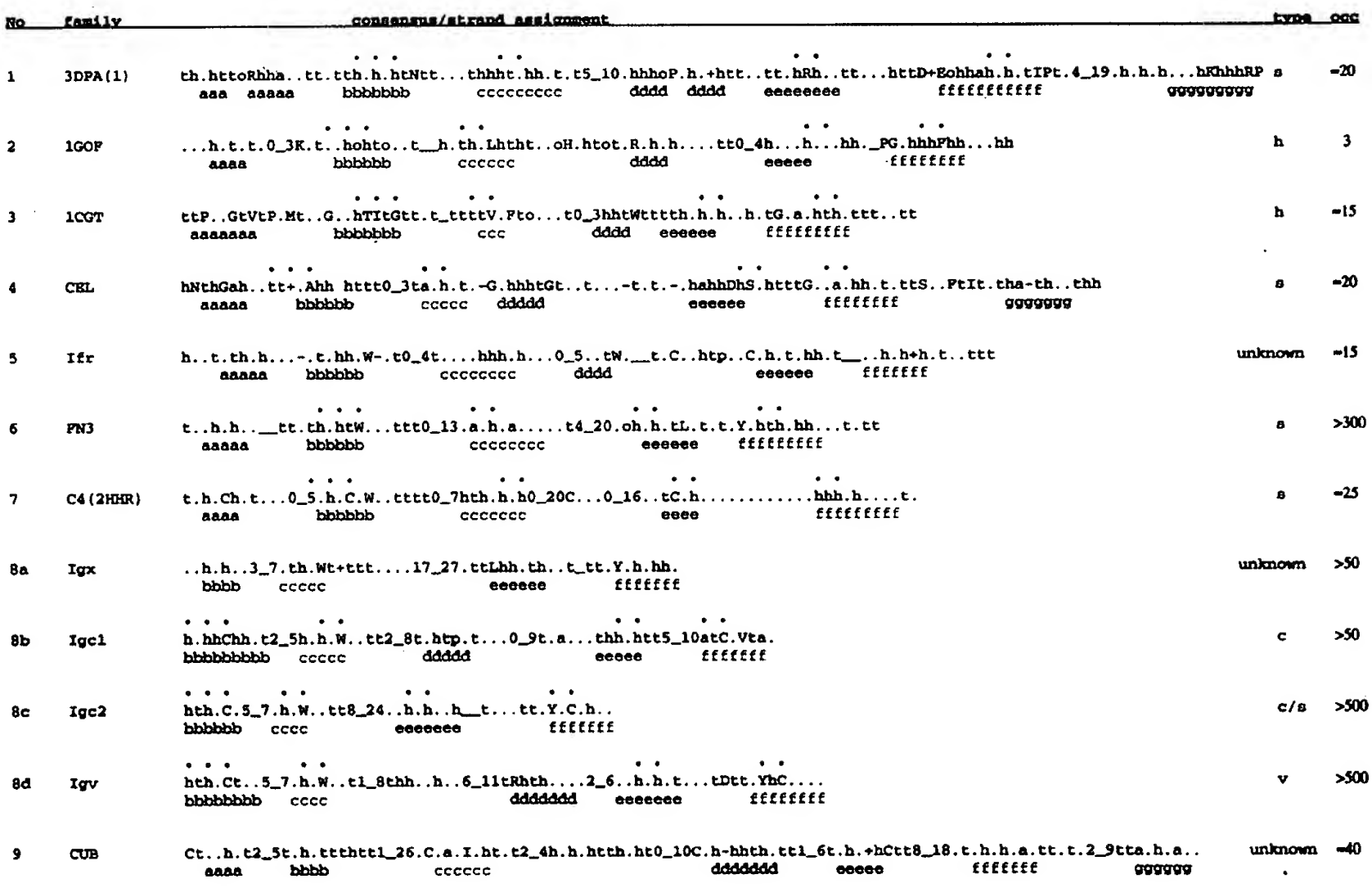
No
family
consensus/strand assignment
type
occ
1
3DPA(1)
th.httoRhha..tt.tth.h.htNtt...thhht.hh.t.t5_10.hhhoP.h.+htt..tt.hRh..tt...httD+Eohhah.h.tIPt.4_19.h.h.h...hKhhhRP
s
=20
2
1GOF
...h.t.t.0_3K.t..hohto..t__h.th.Lhtht..oH.htot.R.h.h....tt0_4h...h...hh._PG.hhhFhh...hh
h
3
3
1CGT
ttP..GtVtP.Mt..G..hTItGtt.t_ttttV.Fto...t0_3hhtWttttth.h.h..h.tG.a.hth.ttt..tt
h
=15
4
CEL
hNthGah..tt+.Ahh htttO_3ta.h.t.-G.hhhtGt..t...-t.t.-.hahhDhS.htttG..a.hh.t.ttS..FtIt.tha-th..thh
s
=20
5
Ifr
h..t.th.h...-.t.hh.W-.t0_4t....hhh.h...0_5..tW.__t.C..htp..C.h.t.hh.t__..h.h+h.t..ttt
unknown
=15
6
FN3
t..h.h..__tt.th.htW...ttt0_13.a.h.a.....t4_20.oh.h.tL.t.t.Y.hth.hh...t.tt
s
>300
7
C4(2HHR)
t.h.Ch.t...0_5.h.C.W..tttt0_7hth.h.h0_20C...0_16..tC.h.....hhh.h....t.
s
=25
8a
Igx
..h.h..3_7.th.Wt+ttt....17_27.ttLhh.th..t_tt.Y.h.hh.
unknown
>50
8b
Igc1
h.hhChh.t2_5h.h.W..tt2_8t.htp.t...0_9t.a...thh.htt5_10atC.Vta.
c
>50
8c
Igc2
hth.C.5_7.h.W..tt8_24..h.h..h__t...tt.Y.C.h..
c/s
>500
8d
Igv
hth.Ct..5_7.h.W..t1_8thh..h..6_11tRhth....2_6..h.h.t...tDtt.YhC....
v
>500
9
CUB
Ct..h.t2_5t.h.ttthtt1_26.C.a.I.ht.t2_4h.h.htth.ht0_10C.h-hhth.tt1_6t.h.+hCtt8_18.t.h.h.a.tt.t.2_9tta.h.a..
unknown
=40

Table 2
Mutation rates of Ig-like domains in vertebrates

Domain	Code	Human *versus* bovine/pig	Human *versus* mouse/rat	Human *versus* chicken	Disulfide bridges
3rd in tenascin	1TEN	96%	89%	86%	No
10th in fibronectin	FN3	93%	86%	83%	No
GHR domain 2	2HHR-2	87%	70%	69%	No
GHR domain 1	2HHR-1	83%	69%	62%	Yes
Microglobulin	3HLA	76%	71%	47%	Yes
CD4 domain 3	1CID-3		59%		No
CD4 domain 4	1CID-4		56%		Yes
CD2 domain 2	CD2-2		56%		Yes
CD4 domain 1	2CD4-1		52%		Yes
CD4 domain 2	2CD4-2		47%		Yes
CD2 domain 1	CD2-1		41%		No
CD8	1CD8-1	56%	41%		Yes

The domains of our dataset (for abbreviations see Table 1) were compared with their putative orthologues from other vertebrates if available. For some of them quantitative comparisons were omitted as too many paralogues exist (e.g. Igs or 2HLA). Human sequences are compared with rodents (mouse, rat), artiodactyla (bovine/pig) and chicken. The sequence identities are averaged if sequences from 2 species of the respective group were found. Fn3 domains appear to be more conserved than other Ig-like domains.

related domains reveal a different location of disulfide bridges within this family as well (data not shown). Structure comparison placed the C4 family in the close neighborhood of another wide-spread sequence family, the Fn3 repeats. The majority of protein domains in this family apparently do not need disulfide bridges at all to stabilize the β-sandwich. Only for a domain in neuroglian has a single unusual disulfide bridge connecting strand *a* and *g* been demonstrated (Huber *et al.*, 1994). However, CD45, which has been predicted to contain Fn3 domains (Bork & Doolittle, 1992), has several cysteine residues in the respective regions and might be a heavily disulfide-bonded example among Fn3 domains.

(d) *Mutation rates*

Each family identified by sequence comparison has clusters of conserved, buried core residues, but these are different between the different sequence families. Because of this strong dispersion of the different families in sequence space, no statement of possible divergence *between* the families can be made. The rate of divergence *within* families can be quantified if orthologues of comparable species have been sequenced. The considerable differences in mutation rates between Ig-like domains in different proteins (Table 2) suggest that functional rather than structural constraints are the dominant influence on the evolution of Ig-like domains. In the examples with

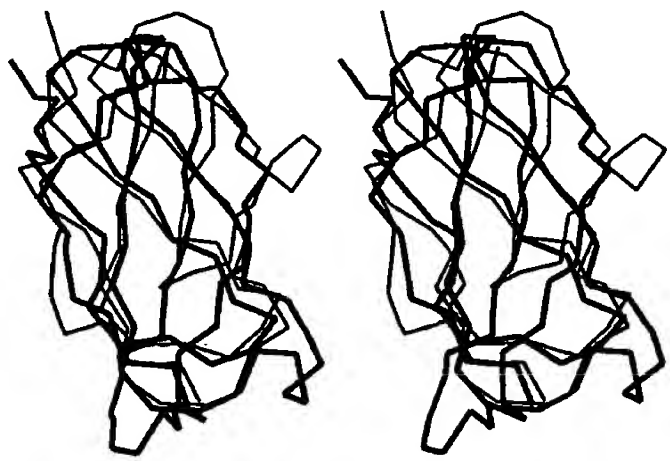

Figure 5. Example of close structural similarity in spite of lack of sequence similarity. Superposition of C^{α} traces: galactose oxidase (bold, 1GOF) and cyclodextrin glycosyltransferase (thin, 1CGT). Sequence pattern of the corresponding families are shown in Figure 4. Structurally equivalent residues (r.m.s.d. of 1·5 Å over 70 C^{α} pairs): in 1GOF residues 543 to 566, 569 to 574, 589 to 596, 600 to 606, 613 to 625, 627 to 638 and in 1CGT residues 497 to 520, 526 to 535, 539 to 542, 546 to 552, 555 to 575, 577 to 580.

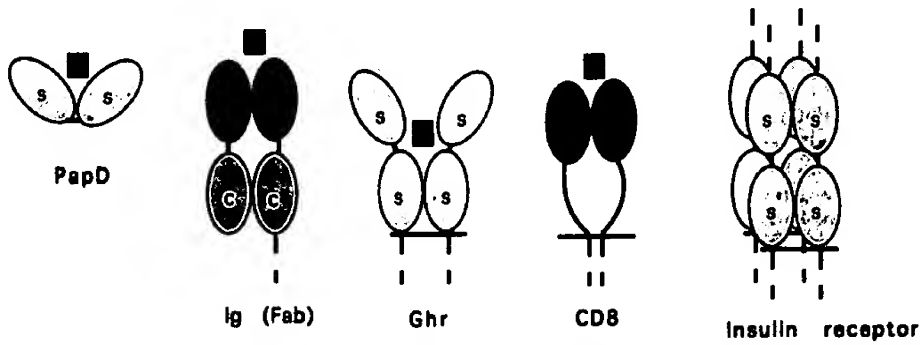

Figure 6. Diversity of binding sites among Ig-like domains. Schema of 5 different modes of binding interaction. Interactions with other Ig-like domains as well as the respective topological subtypes are indicated. Ligands, though different molecules in each case, are depicted as squares.

known three-dimensional structures (Table 2), Fn3 domains have slower mutation rates than domains with sequence similarity to Igs. However, highly conserved domains with sequence similarity to Igs are found in N-CAM neural adhesion molecules (N. Barclay, personal communication): they reach a sequence identity of about 93% between human and rodents (compare with Table 2). Most of the Ig-like domains have, however, a sequence similarity much lower than the average above 90% of all rodent-human protein pairs studied so far (Doolittle, 1987) and thus have faster mutation rates.

4. Unexpectedly Close Structural Resemblances

(a) *Correlation between structure and sequence similarity*

It is well-known that three-dimensional structures are much more conserved in evolution than are sequences. The relationship between structure and sequence similarity is approximately monotonic, i.e. they both decrease in parallel at larger evolutionary distances, down to a threshold level that corresponds to about 25 to 30% identical residues (e.g. Lesk & Chothia, 1986; Doolittle, 1987; Sander & Schneider, 1991; Hilbert *et al.*, 1993). The correlation was verified for the set of Ig-like domains (dataset of Table 1 augmented with additional immunoglobulin domains). Below 25 to 30% sequence identity, however, any correlation is smeared out (data not shown). This can be interpreted in several ways. Either structural dissimilarity, measured as positional r.m.s. deviation, levels off at larger evolutionary distances; or, sequence identity becomes an inadequate measure of sequence similarity below the threshold. Alternatively, convergence in evolution based on physical principles may explain the similarity in structure between some of the very remotely related pairs.

(b) *A putative sugar-binding domain*

In some cases striking structural resemblances are indicative of divergent evolutionary relationships, in spite of the apparent lack of statistically significant sequence similarity. An example is the similarity between Ig-like domains in two apparently unrelated enzymes, cyclodextrin glycosyltransferase and galactose oxidase. Their mutual structural similarity score (see Holm & Sander, 1993) is much higher than that with the other Ig-like domains of our dataset so that they form a distinct subclass of the set of Ig-like domains (Figure 2). The structural similarity extends over the entire domains, except for one loop insertion in galactose oxidase. A total of 70 residues can be superimposed with 1·5 Å r.m.s.d. (Figure 5), a remarkably good agreement at this low level of sequence similarity. A sequence pattern, derived using the structural alignment, clearly discriminates the relatives of the two domains from the random background of unrelated proteins in database searches (data not shown).

The Ig-like domains of cyclodextrin glycosyltransferase and galactose oxidase have a long curled sheet, and topologically they are h-types. Galactose oxidase contains three structurally distinct domains (Ito *et al.*, 1991). The N-terminal domain is apparently a protein module shared with several bacterial sialidases (Bork & Doolittle, 1994). The central domain has a seven-blade propeller fold and contains the catalytically active site and the metal binding site although some residues of the C-terminal Ig-like domain contribute to metal binding (Ito *et al.*, 1991). Cyclodextrin glycosyltransferase also contains distinct structural units. Only domain D has an Ig-like topology. The two calcium binding sites and the active center are located in other regions of the enzyme (Klein & Schulz, 1991). Since many modules of glycosyltransferases are known to bind carbohydrates (Gilkes *et al.*, 1991) and galactose oxidase contains at least one mobile module (Bork & Doolittle 1994) we speculate that the Ig-like module common to both enzymes might also be involved in carbohydrate binding.

(c) *Similarity of growth hormone receptor domain 1 and fibronectin type III repeats*

As already predicted by sequence comparisons (Bazan, 1990; Patthy, 1990) and confirmed by X-ray crystallography (deVos *et al.*, 1992), the second domain of growth hormone receptor is structurally

very similar to Fn3 repeats. The all-against-all-comparison revealed that the first domain, which does not contain the typical Fn3 consensus sequence pattern (Bork & Doolittle, 1992) but which has three disulfide bridges instead, is also structurally more similar to classical Fn3 domains than to other Ig-like domains (Figure 2). Thus it might be a fast evolving domain of the Fn3 type. Interestingly, a tryptophan at the edge of strand *b* is conserved in both families (Figure 4). It is tempting to speculate that tryptophan might be a relic of common ancestry although other common features reduce to some similarity of the hydrophobicity patterns around the β-strands.

5. Comparisons of Binding Features

Ig-like domains occur in functionally extremely diverse proteins. The proteins used in this study (Table 1) represent a rather limited selection yet include functionally diverse matrix proteins, receptors, chaperones and enzymes. They interact with extremely different proteins or ligands varying from small peptides (e.g. HLA) *via* hormones (e.g. GHR) to giant proteins (e.g. titin oligomer). Even more, the determined crystal structures also reveal different binding modes; apparently each part of the surface of the domain can be used for interaction with other molecules (Figure 6).

One common theme is the interaction with other Ig-like domains *via* the sheets. Whereas in the classical Ig variable chains mainly loop regions interact with the ligand, the majority of Ig-like domains appears to interact *via* their β-sheets (Figure 6). Cell surface molecules such as CD4 and CD8 contact domains of MHC class I and class II molecules, CD2 to LFA of other cells. In addition, homo- or heterodimers can be formed. Even within one protein the different structurally similar domains have distinct binding functions. For example, in matrix proteins such as fibronectin or tenascin various binding activities within their Fn3 regions have been reported and the partner molecules range from carbohydrates *via* other Ig-like domains to a network of molecules. The interaction may be mediated by specific regions (e.g. RGD cell surface binding motif in fibronectin) or by parts of the β-sheets. Often two consecutive domains are involved in binding. Examples are growth hormone receptor (deVos *et al.*, 1992), PapD (Holmgren *et al.*, 1992; Slonim *et al.*, 1992) and neuroglian (Huber *et al.*, 1994). The latter binds a metal ion in the cleft between two Fn3 domains (Huber *et al.*, 1994); PapD and growth hormone receptor bind their major substrate in a corresponding region.

Within each sequence family conserved non-hydrophobic residues might hint at the binding mode of a particular family, as has been shown for the PapD-like proteins (Holmgren *et al.*, 1992; Slonim *et al.*, 1992; see Figure 6). However, the members of most other families have distinct binding functions and bind such a diverse set of ligands that only structurally important residues are conserved.

6. Common Fold, Different Folding Pathways?

A comprehensive structural and sequence comparison of currently available Ig-like domains revealed that a common topology of fold is achieved by fundamentally different sequences. The derived sequence consensus lines (Figure 4) do not reveal hydrophobicity patterns common to all Ig-like structures. In some cases a common pattern exists despite a structural divergence (e.g. v-type and c-type), in others (e.g. 2HHR domain 1 and Fn3s) a relatively close structural relationship is not mirrored at the sequence level.

The comparison of all the structures revealed a common structural core composed of two sequence-adjacent pairs of strands (*b-c*, *e-f*). The segment between strands *c* and *e* is extremely variable in sequence. Structurally, the fringes of the common core are extremely variable as shown by variable position, length and number of strands that are attached to the common core. Given the extreme sequence diversity in the set as a whole, no single interaction (or localized set of interactions) can be uniquely identified as a principal determinant of the Ig-like fold. However, the modifications around the common core seen in this study might be of interest to protein engineering. We propose that the topological subclass could change by a single insertion/deletion in the *c'-e* loop region. Disulfide bridges appear to be not essential but might stabilize the fold sufficiently to support high mutation rates (see tendency in Table 2). Disulfides might even mediate drifting of contacts within the core on an evolutionary time scale.

The observed variability opens up intriguing questions concerning the folding pathways of this class of proteins. Because of the non-linear composition of the common core unit, initiation of folding from a single β-hairpin (see Richardson, 1981) appears to be unlikely. However, collapse of (several) locally formed hairpins is somewhat supported since most of the strand-strand interactions in the Ig-like topologies are between sequence-adjacent strands (*a-b*, *f-g*, c-c'-c"). The β-zipper model of Hazes & Hol (1992) proposes initial folding of the *b-c* strands (i.e. a sheet-sheet contact), because the *b-c* loop is usually very short in their sample. In Fn3 repeats, however, the *b-c* loop can contain up to 20 residues. In light of the present data, different folding pathways for different sequence families cannot be excluded. An analysis of the Greek key motif support this view as no single folding pathway is likely to fit all Greek key structures to which the immunoglobulins belong (Hutchinson & Thornton, 1993).

From the variety of features observed in the different Ig-like domains further subtypes and modifications of the topology can be expected. Indeed, the very recently determined structure of cytochrome *f* revealed a modified Fn3-like domain, having additional strands, α-helical elements in the loops and a more barrel-like arrangement (Martinez *et al.*, 1994). What nature can do, protein engineers can mimic. Ig-like domains are a rich source of natural and artificial variation on a single structural theme.

We are grateful to many colleagues for releasing their co-ordinates. In particular we wish to thank Pamela Bjorkman, Bart deVos, Iain Campbell, Harold Erickson, Hazel Holden, Yvonne Jones, Peter Knowles, Robert Poljak and Georg Schulz. We also thank our colleagues from the Protein Design Group at EMBL for software support as well as Rebecca Wade and Neil Barclay for interesting discussions.

References

Barclay, A. N., Birkeland, M. L., Brown, M. H., Beyers, A. D., Davis, S. J., Somoza, C. & Williams, A. F. (1992). *The Leukocyte Antigen Factsbook*. Academic press, New York.

Bazan, J. F. (1990). Structural design and molecular evolution of a cytokine receptor superfamily. *Proc. Nat. Acad. Sci, U.S.A.* **87**, 6934–6938.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyers, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Bork, P. (1993). Hundreds of ankyrin-like repeats in functionally diverse proteins. *Proteins: Struct. Funct. Genet.* **17**, 363–374.

Bork, P. & Beckmann, G. (1993). The CUB domain, a widespread module in developmentally regulated proteins. *J. Mol. Biol.* **231**, 539–545.

Bork, P. & Doolittle, R. F. (1992). Proposed acquisition of an animal protein domain by bacteria. *Proc. Nat. Acad. Sci., U.S.A.* **89**, 8990–8994.

Bork, P. & Doolittle, R. F. (1994). The kelch motif is derived from a common enzyme fold. *J. Mol. Biol.* **236**, 1277–1282.

Brady, R. L., Dodson, E. J., Dodson, G. G., Lange, G., Davis, S. J., Williams, A. F. & Barclay, A. N. (1993). Crystal structure of domains 3 and 4 of rat CD4: Relation to the NH2-terminal domains. *Science*, **260**, 979–983.

Deisenhofer, J. (1981). Crystallographic refinement and atomic models of a human FC fragment and its complex with fragment B of protein A from *Staphylococcys aureus* at 2·9- and 2·8-Angstrom resolution. *Biochemistry*, **20**, 2361–2370.

deVos, A. M., Ultsch, M. & Kossiakoff, A. A. (1992). Human growth hormone and extracellular domain of its receptor: Crystal structure of the complex. *Science*, **255**, 306–312.

Doolittle, R. F. (1987). *On ORFs and URFs*. University science books, Mill Valley, CA.

Doolittle, R. F. & Bork, P. (1993). Evolutionarily mobile modules in proteins. *Sci. Amer.* **269**, 50–56.

Driscoll, P. C., Cyster, J. G., Campbell, I. D. & Williams, A. F. (1991). Structure of domain 1 of rat T lymphocyte CD2 antigen. *Nature (London)*, **353**, 762–765.

Finkelstein, A. V. & Nakamura, H. (1993). Weak points of antiparallel β-sheets. How are they filled up in globular proteins? *Protein Eng.* **6**, 367–372.

Furey, W., Jr, Wang, B. C., Yoo, C. S. & Sax, M. (1983). Structure of a novel Bence-Jones protein (Rhe) fragment at 1·6 Angstroms resultions. *J. Mol. Biol.* **167**, 661–692.

Garrett, T. P. J., Saper, M. A., Bjorkman, P. J., Strominger, J. L. & Wiley, D. J. (1989). Specificity pockets for side chains of peptide antigens in HLA-Aw68. *Nature (London)*, **342**, 692–696.

Garrett, T. P. J., Wang, J., Yan, Y., Liu, J. & Harrison, S. C. (1993). Refinement and analysis of the structure of the first two domains of human CD4. *J. Mol. Biol.* **234**, 763–778.

Gilkes, N. R., Henrissat, B., Kilburn, D. G., Miller, R. C. Jr & Warren, R. A. J. (1991). Domains in microbial 1,4-glycanases: Sequence conservation, function and enzyme families. *Microbiol. Rev.* **55**, 303–315.

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distant similarities. *Proc. Nat. Acad. Sci., U.S.A.* **84**, 4355–4358.

Harpaz, Y. & Chothia, C. (1994). Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. *J. Mol. Biol.* **238**, 528–539.

Hazes, B. & Hol, W. G. J. (1992). Comparison of the hemocyanin β-barrel with other Greek key β-barrels: possible importance of the "β-zipper" in protein structure and folding. *Proteins: Struct. Funct. Genet.* **12**, 278–298.

Herron, J. N., He, X., Mason, M. L., Vos, E. W., Jr & Edmundson, A. B. (1989). Three-dimensional structure of a fluorescein-Fab complex crystallized in 2-methyl-2,4-pentanediol. *Proteins Struct. Funct. Genet.* **5**, 271–280.

Hilbert, M., Böhm, G. & Jaenicke, R. (1993). Structural relationships of homologous proteins as a fundamental principle in homology moidelling. *Proteins: Struct. Funct. Genet.* **17**, 138–151.

Hill, M. O. (1973). Reciprocal averaging: an eigenvector method of ordination. *J. Ecol.* **61**, 237–251.

Holden, H. M., Ito, M., Hartshorne, D. J. & Rayment, I. (1992). X-ray structure determination of telokin, the C-terminal domain of myosin light chain kinase, at 2·8 Å resolution. *J. Mol. Biol.* **227**, 840–851.

Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138.

Holm, L. & Sander, C. (1994). Searching protein structure databases has come of age. *Proteins: Struct. Funct. Genet.* **19**, 165–173.

Holmgren, A. & Bränden, C.-I. (1989). Crystal structure of chaperonine protein PapD reveals an immunoglobulin fold. *Nature (London)*, **342**, 248–251.

Holmgren, A., Kuehn, M. J., Bränden, C.-I. & Hultgren, S. J. (1992). Conserved immunoglobulin-like features in a family of periplasmic pilus chaperones in bacteria. *EMBO J.* **11**, 1617–1622.

Hsu, E. & Steiner, L. A. (1992). Primary structure of immunoglobulins through evolution. *Curr. Opin. Struct. Biol.* **2**, 422–431.

Huber, A. H., Wang, Y. E., Bieber, A. J. & Bjorkman, P. J. (1994). Crystal structure of tandem type III fibronectin domains from drosophila neuroglian at 2·0 Å. *Neuron*, **12**, 717–731.

Hutchinson, E. G. & Thornton, J. M. (1993). The Greek key motif: extraction, classification and analysis. *Protein Eng.* **6**, 223–245.

Ito, N., Phillips, S. E. V., Stevens, C., Ogel, Z. B., McPherson, M. J., Keen, J. N., Yadav, K. D. S. & Knowles, P. F. (1991). Novel thioether bond revealed by a 1·7 Å structure of galactose oxidase. *Nature (London)*, **350**, 87–90.

Jones, E. Y., Davis, S. J., Williams, A. F., Harlos, K. & Stuart, D. I. (1992). Crystal structure at 2·8 Å resolution of a soluble form of the cell adhesion molecule CD2. *Nature (London)*, **360**, 232–239.

Juy, M., Amit, A. G., Alzari, P. M., Poljak, R. J., Claeyssens, M., Beguin, P. & Aubert, J.-P. (1992). Three-dimensional structure of a thermostable bacterial cellulase. *Nature (London)*, **357**, 89–91.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Klein, C. & Schulz, G. E. (1991). Structure of cyclodextrin glycosyltransferase refined at 2·0 Å resolution. *J. Mol. Biol.* **217**, 737–750.

Kraulis, P. (1991). MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950.

Leahy, D. J., Axel, R. & Hendrickson, W. A. (1992*a*). Crystal structure of a soluble form of the human T cell coreceptor CD8 at 2·6 Å resolution. *Cell*, **68**, 1145–1162.

Leahy, D. J., Hendrickson, W. A., Aukhil, I. & Erickson, H. P. (1992*b*). Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science*, **258**, 987–991.

Lesk, A. M. & Chothia, C. (1982). Evolution of proteins formed by β-sheets. II. The core of the immunoglobulin domain. *J. Mol. Biol.* **160**, 325–342.

Lesk, A. M. & Chothia, C. (1986). The relation between divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.

Main, A. L., Harvey, T. S., Baron, M., Boyd, J. & Campbell, I. D. (1992). The three-dimensional structure of the tenth type II module of fibronectin: An insight into RGD-mediated interactions. *Cell*, **71**, 671–678.

Martinez, S. E., Huang, D., Szczepaniak, A., Cramer, W. A. & Smith, J. L. (1994). Crystal structure of cytochrome f reveals a novel cytochrome fold and unexpected heme ligation. *Structure* **2**, 95–105.

Patthy, L. (1990). Homology of a domain of the growth hormone/prolactin receptor family with a typo III module in fibronectin. *Cell*, **61**, 13–14.

Pearson, W. R. & Lipman, D. (1988). Improved tools for biological sequence comparisons. *Proc. Nat. Acad. Sci., U.S.A.* **85**, 2444–2448.

Richardson, J. S. (1981). The anatomy and taxonomy of protein structures. *Advan. Protein Chem.* **34**, 176–339.

Rohde, K. & Bork, P. (1993). A fast, sensitive pattern-matching approach for protein sequences. *Comp. Appl. Biosci,.* **9**, 183–189.

Ryu, S.-E., Kwong, P. D., Truneh, A., Porter, T. G., Arthos, J., Rosenberg, M., Dai, X., Xuong, N.-H., Axel, R., Sweet, R. W. & Hendrickson, W. A. (1990). Crystal structure of an HIV-binding recombinant fragment of human CD4. *Nature (London)*, **348**, 419–426.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.

Saper, M. A., Bjorkman, P. J. & Wiley, D. C. (1991). Refined structure of the human histocompatibility antigen HLA-A2 at 2·6 A resolution. *J. Mol. Biol.* **219**, 277–319.

Slonim, L. N., Pinkner, J. S., Bränden, C.-I. & Hultgren, S. J. (1992). Interactive surface in the PapD chaperonine cleft is conserved in pilus chaperone superfamily and essential in subunit recognition and assembly. *EMBO J.* **11**, 4747–4756.

Vriend, G. (1990). WHAT IF: A molecular modelling and drug design program. *J. Mol. Graph.* **8**, 52–56.

Wang, J. H., Yan, Y. W., Garrett, T. P., Liu, J. H., Rodgers, D. W., Garlick, R. L., Tarr, G. E., Hussain, Y., Reinherz, E. L. & Harrison, S. C. (1990). Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. *Nature (London)*, **348**, 411–418.

Williams, A. F. & Barclay, A. N. (1988). The immunoglobulin superfamily—domains for cell surface recognition. *Annu. Rev. Immunol.* **6**, 381–405.

Edited by I. A. Wilson

(Received 25 March 1994; accepted 17 June 1994)

Note added in proof: The very recently determined three-dimensional structure of *E. coli* β-galactosidase (Jacobson, R. H., Zhang, X.-J., DuBose, R. F. & Matthews, B. W. (1994)). *Nature (London)*, **369**, 761–766 reveals two s-type Ig-like domains that flank the catalytic TIM-barrel.